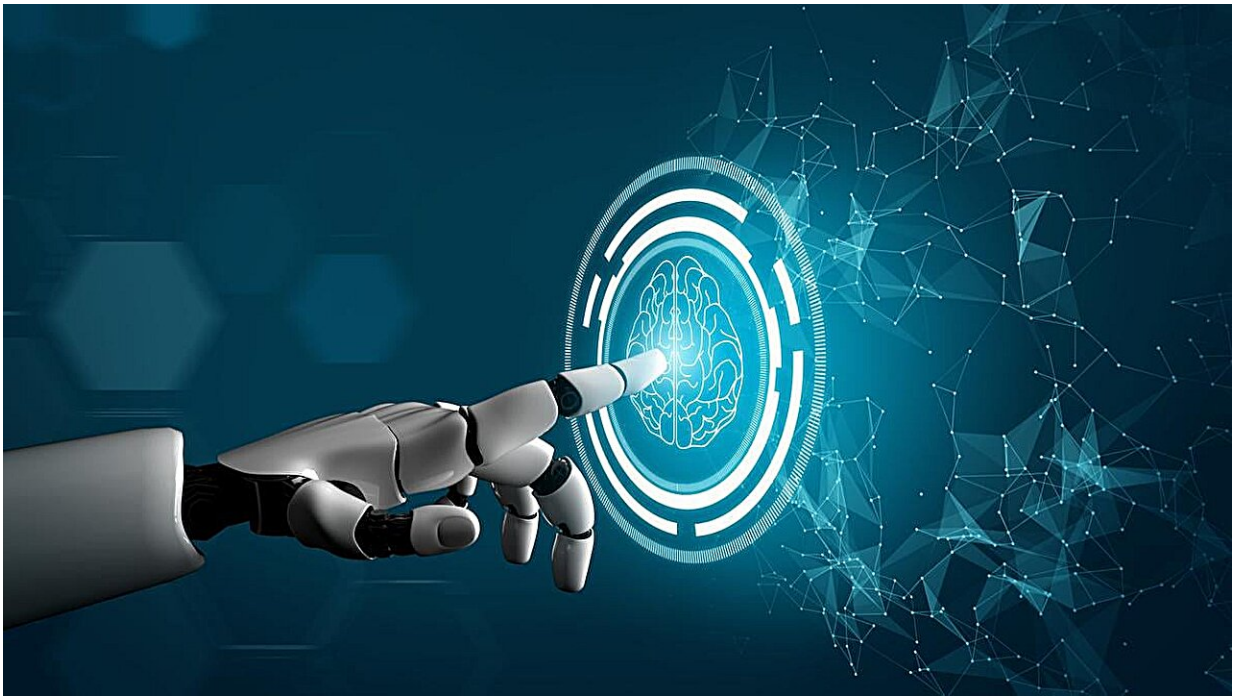
GPT-4 with vision has poor accuracy for image-based radiology questions

September 8 2024, by Elana Gotkine



The large language model GPT-4 with vision (GPT-4V) has high accuracy for text-only radiology questions, but much lower accuracy for image-based questions, according to a study published online Sept. 3 in *Radiology*.

Nolan Hayden, M.D., from Henry Ford Health in Detroit, and colleagues

examined the performance of GPT-4V on radiology in-training examination questions to gauge the model's baseline knowledge in radiology. The September 2023 release of GPT-4V was assessed using 386 retired questions (189 image-based and 197 text-based) from the American College of Radiology Diagnostic Radiology In-Training Examinations; 377 questions were unique.

The researchers found that GPT-4V answered 65.3 percent of the unique questions correctly, with significantly higher accuracy observed on the text-only versus the image-based questions (81.5 versus 47.8 percent). For text-based questions, differences were seen between prompts, with chain-of-thought prompting outperforming long instruction, basic prompting, and the original prompting style by 6.1, 6.8, and 8.9 percent, respectively. For image-based questions, there were no differences seen between prompts.

"We found that while GPT-4V shows relatively good performance on text-based questions, it shows deficits in accurately interpreting key radiologic images. This highlights the model's limitations in visual radiology analysis," the authors write. "We also noted an alarming tendency for GPT-4V to provide correct diagnoses based on incorrect image interpretations, which could have significant clinical implications."

More information: Nolan Hayden et al, Performance of GPT-4 with Vision on Text- and Image-based ACR Diagnostic Radiology In-Training Examination Questions, *Radiology* (2024). DOI: 10.1148/radiol.240153

Francis Deng, Multimodal Models Are Still a Novice at Radiology Vision, *Radiology* (2024). DOI: 10.1148/radiol.242286